2015

Event Detection Using Linear Regression and Historical Weather Data

NICHOLAS GARAFOLA

MAY 2015

Adviser: Dr. Timothy L. Johnson

Masters project submitted in partial fulfillment of the
Requirements for the master of environmental management degree in
The Nicholas School of the Environment of Duke University

Executive Summary

The Research Triangle Institute International (RTI) seeks to minimize its economic and environmental footprint by automating the process of detecting water and energy over- and under- consumption events associated with HVAC systems. The Central Utility Plant, which serves roughly 25 percent of the gross built area on RTI's main campus, has contributed to past overconsumption events due to mechanical failure of cooling tower water makeup float valves. RTI's Facilities Engineering team is looking for a statistical model for real time prediction of cooling tower water makeup consumption in addition to a clear process for updating the statistical model as system parameters change over time.

This project entails development and implementation of a data cleaning tool based in Microsoft Excel and development of a multiple linear regression model using ordinary least squares methodology. The Excel tool includes a set of macros that allow RTI facilities and operations teams to periodically clean new data and update the predictive model in response to changing system parameters that are external to the model, including building uses, HVAC set points and expansions. Analysis of the relationship between historical weather data and cooling tower water makeup consumption entails data cleaning and aggregation and outlier detection. Several linear models explore the relationship between atmospheric temperature, humidity and cooling tower water makeup consumption.

Facilities Engineering would like to implement the equation that yields the most accurate prediction of cooling tower water makeup consumption with RTI's Building Automation System for event detection. An analysis of model problems and out of sample prediction finds that temperature and humidity linear regression models are useful for approximating daily water makeup consumption, but the substantial model residuals indicate that frequent false alarms are probable and that a more rigorous method of analysis is necessary. Additionally, hourly interval models suffer from a low coefficient of determination (R^2), even after accounting for the delayed effect of atmospheric temperature change on demand for chilled water. Confounding variables may cloud the

relationship between hourly changes in atmospheric conditions and cooling tower water makeup consumption.

Based on the imprecise results obtained from linear models, Facilities Engineering may opt to investigate alternative methods for predicting cooling tower water makeup consumption. RTI may also consider system upgrades that would accommodate site collection of atmospheric data for future analyses.

The Excel tool developed for this project allows the Facilities Engineering to clean historical atmospheric data, import consumption data and merge sets to analyze environmental and economic performance of mechanical systems. The Excel tool should be useful for examining relationship between weather and cooling tower water makeup (or energy) consumption for other HVAC systems on the main campus. This project includes a user's guide that explains the functions of the tool as well as methods of troubleshooting and manipulating data.

Acknowledgements

This project was made possible by the support and advice of the following individuals:

Dr. Timothy L. Johnson, Associate Professor of the Practice in Energy and the Environment, Nicholas School of the Environment, Duke University

John Maravich, Master of Environmental Management, Energy and the Environment, Nicholas School of the Environment, Duke University

Jim Miller, Facilities Engineering Manager, RTI International

Gary Bunce, Facilities Engineering, RTI International

Robert D. Helton, Instrumentation Technician, RTI International

Brad Washabaugh, Senior Director, Facility Operations and Planning, RTI International

Dr. Kyle Bradbury, Managing Director, Energy Data Analytics Lab,
Duke University Energy Initiative

Dr. Jesse Daystar, Assistant Director of Corporate Sustainability Programs, Duke Center for Sustainability & Commerce, Duke University

Dr. Steve Sexton, Assistant Professor in the Sanford School of Public Policy, Assistant Professor in the Department of Economics, Duke University

Natalie Olivo, Editorial Consultant, Greater New York City Metropolitan Area

Contact

For more information, please contact: nick.garafola@gmail.com

Contents

Introduction

Research Triangle Institute International (RTI) is an independent nonprofit research institute headquartered in Research Triangle Park (RTP), North Carolina. The RTP campus consists of roughly two-dozen buildings that house offices, laboratories and industrial equipment. Seven laboratory buildings representing roughly 25 percent of the total gross built area on campus obtain chilled water and steam from a central utility plant (CUP).

RTI is interested in minimizing the resource consumption and operating costs associated with campus systems. A recent renovation of the CUP and the implementation of a condensate recovery and return system represent RTI's commitment to improving performance and efficiency with capital investments. Every mechanical system requires routine maintenance and ideally periodic assessment. RTI personnel are challenged with limited time and budget resources and cannot monitor all mechanical components often enough to avoid component issues. The purpose of this project is to build a predictive model that, once implemented, will notify RTI facilities and operations staff via electronic means when a component failure has occurred, based on the computed difference of predicted and observed cooling tower water makeup consumption.

In the spring of 2014, a float valve on one of the CUP's cooling towers failed in the open position, resulting in a continuous stream of water release from the makeup water line. The failure was addressed when RTI's security staff noticed a deluge of water flooding the road adjacent to the CUP and a facilities manager manually inspected the cooling tower. RTI's facilities managers would like to automate the monitoring of cooling tower water consumption by integrating a predictive water use model into the building management system.

Precedent for automated monitoring of campus systems based on meter data exists in RTI's systems. In response to false high readings over hours (or days) from RTI's two sewer meters, facilities employees set up alert thresholds for high flow events in the building management system. If sewer flow rate exceeds a certain volume over a specified interval,

the building management system sends an e-mail alert to facilities managers, who are then able to verify the flow through (or presence of a blockage in) each meter. The success of the sewer outflow billing project indicates that the given threshold usage data, RTI's building management system is capable of issuing automated e-mail alerts.

RTI's Central Utility Plant

RTI's central utility plant serves seven laboratory buildings that lack significant occupancy flux and programmed HVAC setbacks. For purposes of forecasting utility bills and assessing system performance, Facilities Engineering models weekly cooling tower makeup water consumption at the CUP with cooling degree days (CDD). Cooling degree days reflect the difference between the exterior temperature and the balance point, which is the temperature at which facilities require neither heating nor cooling. If the average temperature on a given day is greater than the balance point, the difference between average daily temperature and the balance point represents the number of cooling degree days on the given day. By contrast, if the temperature on a given day is less than the balance point, the difference between the balance point and average daily temperature represents the number of heating degree days on the given day. The facilities served by the CUP are cooling-dominated, which is why Facilities Engineering uses a balance point temperature of 50 degrees Fahrenheit to assess cooling and heating degree days. By contrast, the balance point temperature for residential settings is generally 65 degrees Fahrenheit.

Because some facilities and equipment require cooling at outdoor temperatures less than or equal to 50 degrees Fahrenheit, the building automation system ("BAS") deactivates the water-cooled system and activates the air-cooled chillers (300 tons combined). Based on a visual analysis of historical data of cooling tower water makeup consumption from years 2010 to 2014, cooling tower water makeup consumption is generally zero when the temperature is below 50 degrees Fahrenheit. This relationship is expected due to the very nature of the balance point temperature reflecting the minimum temperature at which

facilities are cooling-dominant. The operation sequence for the chillers validates the lack of positive makeup water consumption values on intervals where outdoor temperature is 50 degrees Fahrenheit or below.

The CUP has three chillers of near-equal capacity: two are 1,300 tons and the third is 1,100 tons. The three cooling towers reject heat from the three chillers. Each cooling tower works by leveraging the energy consumed in the phase change of liquid water. The cooling towers contain spray nozzles and fans which circulate air and water over the closed loop from the chillers. As water makes contact with the hot coils and evaporates, heat is rejected from the closed loop system. When the dry bulb temperature exceeds the wet bulb temperature, the cooling towers are able to reject more heat using evaporative cooling than if they were to rely on air alone.

When a makeup water float valve failure occurs, the float control detaches from the valve assembly and leaves the valve in the open position. As a result, thousands of gallons of water flow freely until the event is manually detected.

Refer to the diagram of the CUP system in the Appendix 1.

Previous Study

Linear regression is a common method for examining related data across a variety of disciplines. Because evaporative cooling is an energy service, a simple linear regression model is a logical first step to examining the drivers of cooling tower water makeup consumption. Although this study relies on linear regression to examine the relationship between temperature, humidity and cooling tower water makeup consumption, a review of past studies indicates that a variety of options of greater complexity are common approaches to modeling the relationship between atmospheric conditions and building energy consumption.

A previous assessment of modeling for chiller energy consumption provides insight on how the model might evolve to account for factors such as wind speed and dehumidification. Lam and others (2009) group days based on prevailing whether conditions and assign the groups to "day type" categories (Lam, Wan, and Cheung 2009). The explanatory variables included dry-bulb temperature, wet-bulb temperature, global solar radiation, clearness index, and wind speed. The methods and results by Lam et al. 2009 are not immediately transferable to the water usage model, in part because the study by Lam et al. considers long-term weather data and utilizes principal component analysis (PCA). The motivation for using PCA instead of ordinary least squares regression is to gain an understanding of the dependencies that exist among explanatory variables. This suggests that correlation among two or more explanatory variables may be a challenge for the water usage regression model.

Although the method of analysis by Lam et al. varies, the high explanatory power of their final model indicates that the explanatory variables (including solar radiation and wind speed) might offer explanatory power in the water usage model. As for the method of analysis, simpler models exist. McMenamin (2008) lists two general approaches for HVAC energy forecasting: rank and average and average by date. The rank by average methods entails assembling a predictive data set based on historical data prior to establishing a function via regression analysis. The first step of the rank by average method entails sorting each year's daily temperature from hottest to coldest. McMenamin provides an example in which the average, minimum and maximum values for each year are ranked independently from each other. The first example ranks data within each historical month and then computes the average daily temperature across each month. Average monthly values are then computed over a number of years. The result is a relatively smooth, downward-sloping set of minimum, average and maximum curves on a plot of temperature over percentile ranking.

The nature of data collection and analysis in this study more closely resembles methods for establishing a weather-normalized baseline in the context of measuring and verifying discrete energy efficiency improvements. The two-stage method entails collecting and

analyzing historical data on consumption, and then evaluating consumption data that occur after an energy efficiency upgrade or system change (Agnew & Goldberg,, 2013). This analysis aims to establish the historical consumption baseline once in the form of a linear equation. The primary difference is that this project focuses on an equation that will be evaluated regularly for each established interval of operation, resulting in a predicted value that the BAS can compare to the interval consumption value. The same method, linear regression, can be used to interpret consumption trends for whole-building retrofits or for event detection. The unique approach to this project is the repeated and regular use of the linear model to predict values.

Data

This analysis relies on data from two distinct sources: historical weather data from a third party and cooling tower water makeup consumption data from RTI's building automation system (BAS). Although RTI collects temperature data at 15-minute intervals on-site, insufficient data were available at the time of this analysis.

Prior analyses exploring the relationship between atmospheric conditions and cooling tower water makeup consumption relied on daily interval data. Because this analysis must account for much more precise intervals (hourly or intra-hourly), data collected must be as granular as is economically feasible. An evaluation of data from the National Oceanic and Atmospheric Administration ("NOAA") for this study found that the hourly interval data is the most precise interval data available at no cost to the user (The National Oceanic and Atmospheric Administration (NOAA), 2015).

The weather data of choice is quality-controlled local climatological data from the NOAA. The data is available on an hourly interval basis from NOAA's National Climatic Data Center (The National Oceanic and Atmospheric Administration (NOAA), 2015). NOAA data are collected at the Raleigh/Durham International Airport Station, which offers daily and hourly data for each month. The monthly data sets are available in HTML and ASCII (files

bearing the extension ".CSV") formats. Although ASCII format is native to Microsoft Excel, the Excel tool has been built with the assumption that the user will view each month's data in HTML format and paste it into the corresponding month's sheet in the Excel tool. This is not only the client's preference based on the formatting of the HTML data but facilitates user review of the data prior to analysis.

RTI Data Overview

This analysis is based upon cooling tower water makeup consumption data export in six-month sets for the 2013 and 2014 calendar years. The Excel tool is built to handle data sets that include hourly observations for multiple variables, including cooling tower blowdown, makeup water consumption, condensate recovery and sewer meter 3, which is the sewer outflow that captures cooling tower water overflow.

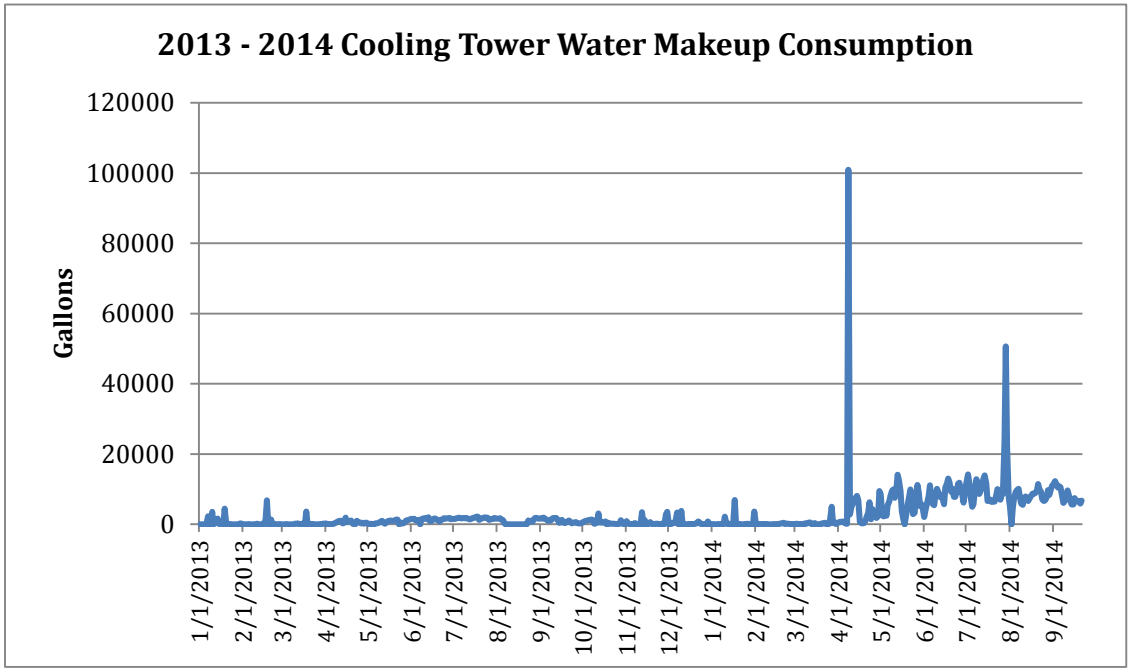
Systems data are collected by mechanical meters connected to RTI's building automation system, which communicates with meters, adjusts HVAC system controls, stores data on a server and makes the data accessible to Facilities Engineering.

The backbone of RTI's BAS is the R2 platform, which interfaces Ax, a successor. Ax is what allows RTI's data visualization platform, Periscope by Activelogix, to communicate with the R2-based system. System sequences and settings are still imported into a back-end Java interface in R2, whereas visualization of data occur within Periscope's web-based graphical user interface (GUI).

The sequence of communication between the systems may be responsible for occasional extreme outliers and other types of invalid values in the BAS data. Cooling tower water makeup consumption readings reflect the total volume of water (gallons) consumed over each interval, so negative values are considered erroneous. Positive extreme outliers also exist. Data for modeling expected values of cooling tower water makeup consumption must exclude negative values and extreme outliers for reasons explored below.

The historical data include a range of values deemed "normal" by Facilities Engineering and extreme positive outliers, which represent a combination of actual high consumption events and erroneous readings. Figure 1 shows data that have been cleaned manually by Facilities Engineering to exclude negative values but not outliers.

Figure 1: Historical (Cleaned) Consumption Data Provided by RTI Facilities Engineering



High readings that occur in April, July and August 2014 reflect measurements of water consumption during float valve failure events. The data for this analysis will have to exclude outlier values in an effort to establish a relatively narrow linear trend that predicts expected water consumption and provides a robust model for detecting outlier-range consumption.

Of concern is the apparent difference in magnitude between the 2013 and 2014 water consumption data. The 2013 data appear to be an order of magnitude lower than the 2014 data. The greater overall magnitude of the 2014 data is due in part the float valve failure events previously described, but the high magnitude of the non-event readings in 2014 relative to those from 2013 indicate a major system change between cooling seasons or a metering error.

Metering errors may occur due to mechanical malfunction (i.e. jamming) or electronic errors, such as a pulse multiplier problem. A pulse multiplier problem occurs when a meter's pulse multiplier, which provides the relationship between the volume measured by the meter and a numeric reading to the BAS, is mistakenly set too low or high. The resulting data are highly correlated with the true consumption data but vary by a consistent percentage. Regardless of the type of metering error, the 2014 data suggest that the cooling tower water makeup meter was adjusted prior the start of the cooling season.

NOAA Data

This study relies on hourly interval time series data obtained in monthly sets from the National Oceanic and Atmospheric Administration (NOAA) Quality Controlled Local Climatological portal (The National Oceanic and Atmospheric Administration (NOAA), 2015). The quality-controlled NOAA data is vertical time-series data, meaning that the column headers represent the names of the variables observed, including temperature, humidity, and dew point. Each row of data has unique day and time values in the first two columns. Subsequent columns reflect the value of observations occurring at each time interval. Despite quality control measures, monthly data are subject to both missing and extra intervals. The remainder of this section details some of the barriers associated with using NOAA data.

NOAA's quality-controlled data is accepted by RTI's facilities management team with the caveat that the data must be cleaned prior to use in analyses. The amount of time associated with cleaning the data constitute a portion of the motivation for this study, as the facilities management team seeks to find way to obtain clean weather data routinely and efficiently.

The foremost task required for structuring an analysis based on NOAA quality-controlled data is assembly of data from individual months. NOAA's website (http://cdo.ncdc.noaa.gov/qclcd/QCLCD?prior=N) directs the user to select a location,

month and interval for viewing. This results in a one-month data set without year and month information included in the variables. The first two columns of NOAA data include variables for day (integer values ranging from 1 – 31) and time.

Time intervals in NOAA data are expressed differently from the timestamp format in RTI's BAS system and the format recognized by Microsoft Excel. The regular observations occur on the 51st minute of every hour in the format #:51, so the minute counts must be converted into hours. When missing observations do occur, the variables assume non-numeric values for the unobserved interval (e.g. "M" for missing). Equally problematic are extra observations, or those that occur in between each 51st interval. These observations sometimes occur as duplicates (i.e. two observations with the same time stamp), but they also occur at other time intervals.

The Building Automation System data have very different characteristics and therefore require very different cleaning procedures than those used for processing the NOAA data. For the purposes of this analysis, the BAS data are .CSV exports from the Periscope Graphical User Interface and consist of hourly volume observations of cooling tower makeup consumption,

In addition to outlier readings, the Periscope exports are subject to transcription errors due in part to translation between the underlying software platform (R2 and AX) and the Periscope interface.

Past Models

RTI's Facilities Engineering team used daily interval cooling degree days (CDD) from Weather Underground to examine the relationship between demand for cooling services (based on outdoor temperature) and cooling tower water makeup consumption. Facilities Engineering hypothesized that the relationship between weekly cooling degree days and cooling tower water makeup consumption changed in April 2014,

Figure 2: Overview of Historical Makeup Water Consumption Data and Temperature

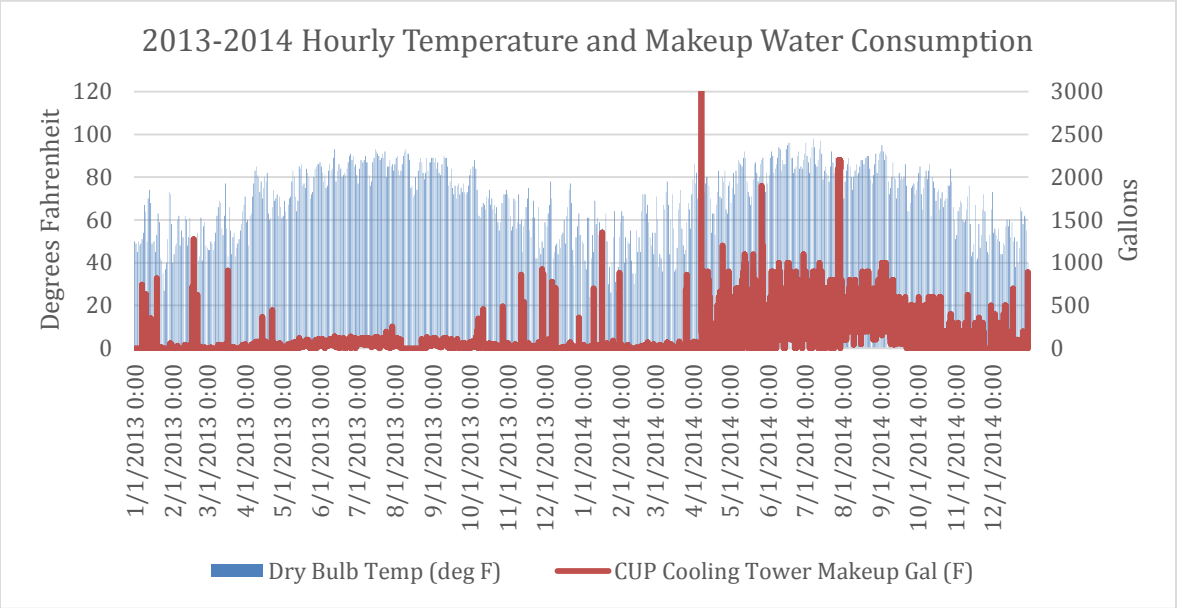
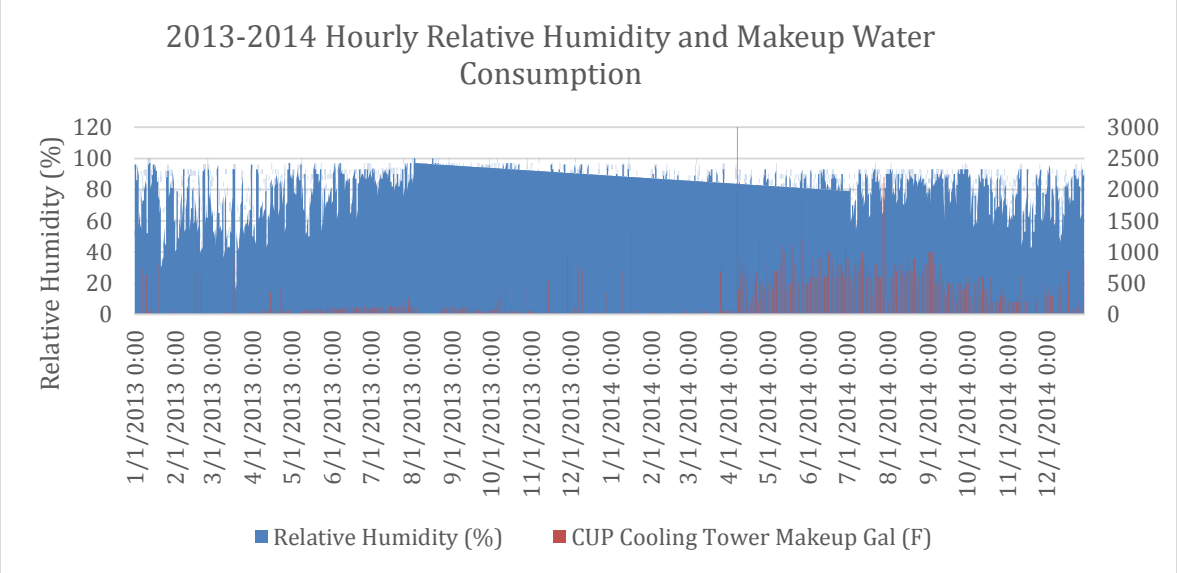


Figure 3: Overview of Historical Makeup Water Consumption Data and Relative Humidity



around the time facilities and operations teams were incorporating a condensate recovery system that was expected to reduce the amount of cooling tower water makeup consumption across all temperature values. S**e**parate linear models examine the relationship before and after the major spike in water consumption (early April 2014). The analysis finds two similar, but unique trends resulting in substantial differences in annual

consumption ($R^2 > .90$). The results are likely due to equipment malfunctions and not the addition of the condensate recovery system.

Methods

The data cleaning, alignment and variable computation processes are done with macros in Excel that automate the process for each calendar year of data. This section describes the processes undertaken by a series of macros and the subsequent manual steps (aggregating hour intervals to daily intervals).

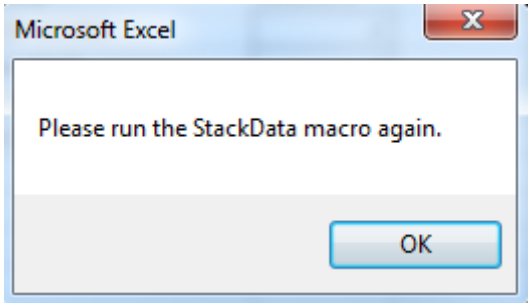
General Coding Practices

Code for this project has been developed largely by manually performing data cleaning and analysis in Excel prior to developing subroutines. A subroutine or "macro" in this context is a set of procedures written in Microsoft Visual Basic for Applications ("VBA") to handle the automation of repeatable tasks in Excel 2013.

Coding occurs with the assumption that the client will use the Excel tool to clean and analyze future data sets with the same general layout and variables. The user inputs on the CONTROL sheet and variable row and column references allow users to alter the tool such that they may explore the relationship between weather data and alternative dependent variables (e.g. energy consumption). However, the weather data cleaning macros are unique to data problems within NOAA (particularly time interval transformation) and will not work with weather data from other sources.

An objective of the project is to deliver an Excel tool that can grow with RTI's needs and be serviceable by future interns who have some familiarity with VBA. The following best practices have been applied throughout the VBA macros that constitute the backbone of the Excel tool:

- **Annotation**. Annotations explain the purpose and function of each block of code, in addition to explanations above each formula and loop. Annotations are added to define the objectives of subroutines and often individual lines of code.
- **Indentation and Spacing**. In addition to the section headers and brief explanations, indentations and blank lines allow the editor to easily identify individual blocks of code.
- **Minimal loops**. Most macros in the Excel tool use filters or assign formulas to ranges to avoid massive loops. This allows the Excel tool to quickly evaluate criteria based on built-in Excel functions. Looping through every single data increases run time and contributes to stability issues. Large loops have been minimized to ensure that the Excel tool is effective and efficient.
- **Paste values**. The Excel tool subroutines use formulas across ranges to calculate values. Once the formula has been applied to the entire range, the calculated values are pasted over the formulas to maintain a relatively small file size.
- **Automatically assess variable ranges**. Count rows and columns instead of coding static assumptions or prompting the user for each subroutine.
- **Testing**. Run macros with variable data sets to ensure that reference variables and assumptions have been accurately coded to ensure that macros can handle subsequent data.
- **Error checking.** If the required data are not present in a macro, the macro will yield error messages which are often vague. In order to help the user understand errors, each macro contains small tests to see if the requisite data are present prior to computing values. If the necessary data are not present in the Excel tool, the macro warns the user with a specific warning that has been hard-coded into each macro. An example is:



Error messages can be tested by running each macro (aside from the first) without running the macro listed prior.

Debugging Macro Code

Debugging VBA macros in Excel entails stepping through each line of code and the spreadsheet as well as the values each variable assumes to determine the functionality of the macro and correct or enhance the code. In the example below, the macro is attempting to add a filter to a table in the range of A1:L8763 in order to exclude negative cooling values for hourly cooing tower water makeup consumption. When Visual Basic is in Break Mode, the code below shows a selection class error on line 2 and highlights the number 10. This indicates that the value 10 is not a valid input. The debugging shows that variables FilterRow and LastRow are correctly reflecting 1 and 8763, respectively. Upon closer inspection, the "A" highlighted in Line 1 should not be equal to A but to L, which is the unique last column of the table based on the size of this data set. Upon changing the highlighted A to the variable representing the letter of the last column, the code functions properly.

```
Line 1
Set RFilter = Range("A" & FilterRow & ":" & "A" & LastRow)
Line 2
Set rr = Range(MakeupColLet & FirstRow & ":" & MakeupColLet &
_LastRow)
Line 3
RFilter.AutoFilter Field:=10, Criteria1:="<0",
Operator:=xlFilterValues
```

Assembling and Cleaning Data

Because NOAA data are captured as individual months and lack year and month variables, the first macro, named "StackData," creates new columns, computes month and year values, and stacks months to form a year's data set. The order of the data cleaning and merging macros has changed several times throughout the coding. Assembly of weather data and BAS data occur prior to cleaning because a component of cleaning is ensuring that hourly

intervals are consecutive. Likewise, the Excel tool has been developed in such a way that each data set is cleaned prior to prior to merging. Merging cleaned data sets is quicker than merging data sets with invalid time intervals because the data sets are shorter once they are clean.

The first macro called from the user control panel is StackData, which looks at data from each of twelve month tabs and copies the data to a sheet called "Master." The macro begins by testing to see if sheet 'Master' exists. If the sheet doesn't exist, StackData creates it. StackData then adds two columns to each month's sheet: column A becomes the year value in the format "yyyy", while column B becomes the name of the month. The values for year and month name are applied to every row of the individual month sheets as well as the combined data set. The existing variable columns are pushed to the right, extending the width of the tables. Built-in logical tests check to ensure that month and year columns cannot be recreated if the user accidentally runs StackData more than once, as having multiple sets year and month columns would further shift the subsequent variable columns and skew cell references for subsequent macros.

After combining the NOAA data into one large set, StackData converts time values from the "##51" format integer values 0 – 23 to maintain consistency with the 24-hour time format used in the BAS data from Periscope. The general formula used in the macro is:

$$Hour = (Original\ Value - 51) / 100$$

For extra intervals that occur in between regular measurements (i.e. timestamps that do not end in "51"), the formula yields a decimal value. Intervals with decimal time values are deleted by the data cleaning macro (described in the next section) because the extra intervals do not align with the BAS observations, which occur hourly.

StackData calls a macro named "DealWithTime" to add a series of time variables that are used by the final macro to merge the NOAA and BAS data sets. DealWithTime creates five additional columns to the left of the year and month column previously inserted by

StackData. The five time variables result in a daylight-saving-adjusted time value a subsequent macro uses to align the NOAA and BAS data sets. The multiple variables also provide the user options for comparing the data with other sets.

DealWithTime adds the following time variables:

- Full Date. This variable extracts month, day and year from the first three variables of the data to yield the date in the format "MM/DD/YYYY."
- Time in "#:##" format.
- Serial Time.
- Serial Date.
- Serial Date : Serial Time. (this is the variable that unifies the two data sets)

Once the NOAA data have been stacked, the user is able to press the "Transform" button on the control panel, which calls the macro "Transform." Transform deletes duplicate observations as well as those that do not occur on hourly intervals, formats relative humidity as percentile values ("0%"), and calls the macro that calculates cooling and heating degree hours as well as the macro that eliminate intervals with invalid cooling and heating degree hour values (due to missing or non-numeric dry bulb temperature data). Based on logic adapted from calculating cooling and heating degree days (Deliso, 2013), cooling degree hours and heating degree hours are computed with the following logical statements:

- If the average hourly temperature is below 50 degrees Fahrenheit, compute Heating Degree Hours = 50 – (average hourly temp). Otherwise, Heating Degree Days = 0.
- If the average hourly temperature is above 50 degrees Fahrenheit, compute Cooling Degree Hours = (average hourly temp – 50). Otherwise, Cooling Degree Days = 0.

The weather data cleaning tool allows the user to export clean weather data independently of running a regression analysis for cooling tower water makeup consumption.

Aligning Data

The Excel tool aligns the BAS and NOAA data into one cohesive set based on the "Serial Date : Serial Time" value established by DealWithTime and CleanBAS (the macro that cleans the BAS data). The final macro, AlignSets, ensures that the only the NOAA data intervals corresponding with each BAS interval are inserted into the merged set. AlignSets starts with the clean BAS data and applies VLOOKUP formulas for each of the NOAA variables, using the "Serial Date : Serial Time" value as the lookup value in each formula. Although this process corrects for misalignment due to differences in Daylight Saving Time between the data sets, it initially yields error values for intervals in which BAS data are present and NOAA data are missing.

The solution in AlignSets is similar to a process used in the NOAA and BAS cleaning macros: filter and delete unacceptable values. In previous macros, the filters applied to non-numeric, negative and outlier values. In AlignSets, the filter reveals all rows marked "#N/A" as a result of missing NOAA intervals. The macro then deletes the error values and removes the filter from the spreadsheet. As is the case with the earlier macros, much of the code in AlignSets addresses the variable numbers of rows and columns in the data, in addition to formatting and filtering.

Process for Using the Excel Tool

The concept of the models for predictive analysis of cooling tower water consumption at RTI is as follows:

1. User pastes individual monthly NOAA data into the corresponding month sheets within the Excel tool.
2. User presses "Stack Data" button. NOAA data are stacked (individual months feed into year-long data set).

3. User presses “Transform” button. NOAA data are cleaned (non-consecutive, duplicate and non-numeric intervals are removed); time, date and lagged temperature variables are computed.
4. User presses “Import” button and imports BAS data by selecting .CSV files previously exported from Periscope.
5. User presses “Stack BAS” button. BAS data are stacked (individual files feed into one data set).
6. User presses “Clean BAS” button. BAS data are cleaned (sample mean and standard deviation are computed and outliers existing outside the user-defined threshold are removed). Time and date variable columns are created and computed.
7. User presses “Align Data” button. NOAA weather data and BAS data feed into a merged data set based on the time and date serial number from each NOAA observation that matches the time and date serial number from each BAS observation (this process controls for changes in DST).
8. User presses the “Export Set” button to export the merged data set.

Linear regression models use clean versions of the final combined data set from the Excel tool. The amount of time required to operate the Excel tool to prepare clean data is a small fraction of the time required for assembling, cleaning and merging the data sets manually.

Models

Cooling tower water makeup consumption is driven by temperature and humidity both directly and indirectly. Demand for chilled water from the CUP is a function of outdoor temperature and humidity, which in turn contributes to cooling tower water makeup needs. Cooling tower water evaporation and the makeup consumption that replaces evaporative loss is also a function of the ambient temperature and humidity conditions surrounding the cooling towers. The combined data sets produced by the Excel tool offer multiple measures for temperature and humidity, including:

- Dry Bulb Temperature
- Wet Bulb Temperature
- Relative Humidity
- Dew Point Temperature
- Cooling Degree Hours

The challenge with linear regression models is that for a model to yield accurate predictions, several assumptions must hold true. The first is the assumption that the model residuals, or the distances between each data point and the prediction offered by the linear model, are uncorrelated with explanatory variables. Residuals should be normally distributed around zero (meaning that residuals further from zero are less likely to occur). Also, the explanatory variables need to be uncorrelated with each other in order to obtain the clearest relationship between each predictor and the dependent variable (Boslaugh, 2012).

The initial regression models follow certain criteria in part due to weather patterns, system changes (or errors) and uncertainty. The first criterion is year. Because the 2013 cooling tower water consumption data appear to be an order of magnitude lower than the 2014 data, initial models assess the relationship between data from each year independently. The downside associated with using fewer observations in any statistical analysis is a shorter amount of intervals over which the trend is computed. The result is greater uncertainty over the difference between the predicted values from the model and the true values. For the purposes of comparison, subsequent models use the combined data from 2013 and 2014.

The second criterion for the initial linear regression models is cooling season status. During the winter months, the majority of observed intervals have outdoor dry bulb temperatures equal to 50 degrees Fahrenheit or less (described here as "low outdoor temperature(s)"). Low outdoor temperatures, per Facilities Engineering, signals the C.U.P. to meet demand for chilled water with air-cooled chillers. At temperatures below 50

degrees Fahrenheit, cooling towers are not in operation. The lack of operation of the cooling towers during the majority of intervals in December, January and February results in makeup water consumption values equal to zero. The presence of zeroes in the model does not describe the system behavior during the cooling season. Initial models are based on data that exclude observations where makeup water consumption values equal zero in order to better assess the relationships that exist between positive cooling degree hour (or day) values and positive cooling tower water makeup consumption.

The third and final criterion for the initial models is the interval, or granularity. Previous analyses conducted by Facilities Engineering used weekly interval data. The data from NOAA and therefore the data from the Excel tool occur at hourly intervals, permitting a more granular analysis.

Initial Base Models

In order to best understand the drivers of cooling tower water makeup consumption, this analysis uses multiple models with overlapping characteristics and compares their results. In particular, the initial base models examine two time intervals, hourly and daily. Models 1-2 examine daily intervals (the top half of Table 1) due in part to low R^2 values from hourly models (the bottom half of Table 1) and to provide Facilities Engineering with the option to implement a daily model in the BAS instead of the hourly model. The first set of models (Tables 1 - 3) uses cooling degree hours (CDH) calculated from dry bulb temperature and relative humidity (percent) as explanatory variables. The general equation representing each of the initial models is:

*Makeup Consumption = Beta naught + Beta1*CDD 50 + Beta2*relative humidity*

Table 1: Initial Regression Models 1-4

Dependent Variable : Cooling Tower Water Makeup Consumption (Gal)

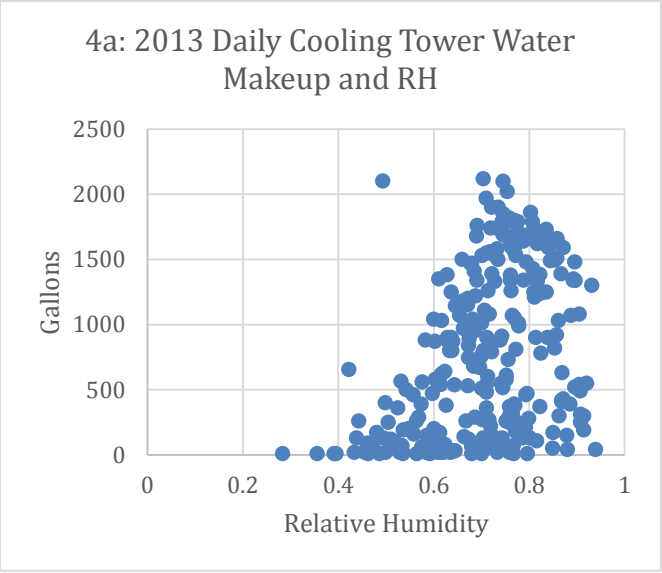
Models use observations where the dependent variable > 0

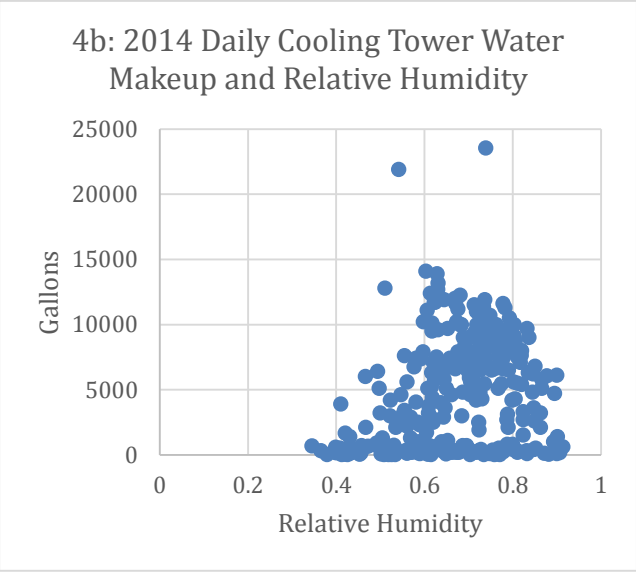
	2013			2014		
	Intercept	RelHum	CDD, CDH	Intercept	RelHum	CDD, CDH
Interval: Daily						
Model 1: Full Year	-462.186	3.494	61.072	-1758.401	5.842	374.781
Std Error	94.388	1.422	1.762	614.604	9.291	11.654
Alpha Level		0.01	0.05		0.25	0.01
t stat	-4.897	2.456	34.670	-2.861	0.629	32.160
n		219			288	
R-Square		88%			80%	
Correlation Residuals : D.V.		0.00%	0.00%		0.00%	0.00%
Model 2: April 1 - October 31	-879.972	7.463	66.850	-3566.793	18.267	418.009
Std Error	111.505	1.598	2.000	924.716	12.918	18.574
Alpha Level		0.01	0.01		0.10	0.01
t stat	-7.892	4.671	33.424	-3.857	1.414	22.505
n		174			210	
R-Square		89%			72%	
Correlation Residuals : D.V.		0.00%	31.33%		0.00%	0.00%
Interval: Hourly						
Model 3: Full Year	-17.399	0.259	2.181	-33.015	0.811	13.432
Std Error	2.854	0.029	0.061	16.798	0.185	0.372
Alpha Level		0.01	0.01		0.01	0.01
t stat	-6.096	9.066	35.747	-1.965	4.396	36.148
n		4166			4466	
R-Square		24%			23%	
Correlation Residuals : D.V.		0.00%	0.00%		0.00%	0.00%
Model 4: April 1 - October 31	-37.156	0.417	2.540	13.579	0.423	12.898
Std Error	2.672	0.025	0.054	23.772	0.231	0.484
Alpha Level		0.01	0.01		0.05	0.01
t stat	-13.904	16.419	47.364	0.571	1.827	26.663
n		3735			4012	
R-Square		38%			17%	
Correlation Residuals : D.V.		0.00%	0.00%		0.00%	0.00%

Because relative humidity reflects the amount of moisture in the air in relation to the maximum level of humidity the air can hold at a given temperature, the observations for relative humidity should not be correlated with dry bulb temperature and therefore cooling degree hours. The initial models offer several encouraging observations. First, the estimated coefficients for temperature and humidity have near-consistent statistical significance across Models 1-4. The notable exception is Model 1, in which the alpha level for relative humidity is 25 percent. The alpha level is important because it reflects the probability of committing a Type I error, or the chance or finding a relationship between variables when a relationship does not really exist (Boslaugh, 2012).

One reason for lower than desired statistical significance (reflected in an alpha level equal to 0.25) for the estimated coefficient of relative humidity in Model 1 (2014) is that the relationship between cooling tower water makeup consumption and relative humidity is non-linear. The decidedly non-linear relationship between the variables (Figure 4a-b) results in higher model residuals, a lower model R^2 and ultimately imprecise predictions.

Figure 4a-b: Scatterplots of Water Makeup Consumption and Relative Humidity





Nested Models

The poor functional relationship between relative humidity and cooling tower water makeup consumption necessitates a comparison of models with and without relative humidity as an independent variable. Models 5-8 (Table 2) use single linear regression to examine the relationship between temperature (CDH/CDD) and cooling tower water makeup consumption, intentionally omitting relative humidity. The comparison of the full to nested models (Table 1 and Table 2) indicates whether including relative humidity as an explanatory variable results in a less precise model. The estimated coefficients for temperature should have consistent signs between the full models (1-4) and the respective nested counterparts (5-8). If including relative humidity as an explanatory variable detracts from the model, the R^2 value should increase and residuals should decrease.

A visual comparison reveals that regression Models 5-8 do not differ significantly from Models 1-4. The signs and magnitudes of the estimated coefficients for temperature are consistent and the measures remain statistically significant. In some cases, the R^2 value is slightly lower, but only by one or two percentage points, with the exception of Model 8, which is five full points lower than the R^2 of Model 4. Additionally, the hourly models (3, 7, 4, 8) all have R^2 values that are much lower than expected (values greater than 50 percent are common for time series data). The analysis section explores potential causes and remedies for the lack of determination by the hourly models.

Table 2: Initial Models, Omitting Relative Humidity

D.V. : Cooling Tower Water Makeup Consumption (Gal)

Models use observations where the dependent variable > 0

	2013		2014	
	Intercept	CDD, CDH	Intercept	CDD, CDH
Interval: Daily				
Model 5: Full Year	-243.926	62.855	-1396.797	376.659
Std Error	32.195	1.623	216.735	11.253
Alpha Level		0.01		0.01
t stat	-7.577	38.718	-6.445	33.472
n	219		288	
R-Square	87%		80%	
Correlation Residuals : D.V.		0.00%		0.00%
Model 6: April 1 - October 31	-396.283	69.777	-2393.980	422.067
Std Error	43.770	2.011	409.955	18.395
Alpha Level		0.01		0.01
t stat	-9.054	34.698	-5.840	22.945
n	174		210	
R-Square	87%		72%	
Correlation Residuals : D.V.		0.00%		0.00%
Interval: Hourly				
Model 7: Full Year	5.473	1.995	30.045	13.046
Std Error	1.348	0.058	8.759	0.362
Alpha Level		0.01		0.01
t stat	4.062	34.385	3.430	36.059
n	4166		4466	
R-Square	22%		23%	
Correlation Residuals : D.V.		0.00%		0.00%
Model 8: April 1 - October 31	2.485	2.139	52.219	12.493
Std Error	1.186	0.049	10.859	0.430
Alpha Level		0.01		0.01
t stat	2.096	43.271	4.809	29.045
n	3735		4012	
R-Square	33%		17%	
Correlation Residuals : D.V.		0.00%		0.00%

The motivation for combining 2013 and 2014 daily data is to increase the number of observations (n). Although the 2013 and 2014 data are subject to system changes, including the removal of a building and the addition of a condensate recovery system, the greater number of observations *n* may illustrate a clearer relationship relative to a model based off one year of data alone. The results in Table 3 do not yield much higher R^2 values than the previous models, although both estimated coefficients are statistically significant at the one percent level in Model 10. Increasing the number of observations does not greatly improve the initial models.

Table 3: 2013 – 2014 Combined Models

D.V. : Cooling Tower Water Makeup Consumption (Gal)

Models use observations where the dependent variable > 0

	2013 - 2014 Combined		
	Intercept	RelHum	CDD, CDH
Interval: Daily			
Model 9: Full Year	-2067.856	10.392	372.439
Std Error	427.334	6.436	8.141
Alpha Level		0.10	0.05
t stat	-4.839	1.615	45.751
n	507		
R-Square	83%		
Correlation Residuals : D.V.		0.00%	0.00%
Model 10: April 1 - October 31	-4079.531	26.149	413.086
Std Error	596.262	8.377	11.562
Alpha Level		0.01	0.01
t stat	-6.842	3.121	35.727
n	384		
R-Square	79%		
Correlation Residuals : D.V.		0.00%	0.00%

Analysis

The ideal criteria for each linear model include statistical significance at or below the 5 percent level, residuals that are normally distributed and uncorrelated with explanatory variables, and an R^2 value of 90 percent or greater. The statistical significance of relative humidity in several of the initial models is above the 5% level, which may be due to the lack of correlation between cooling tower water makeup consumption and relative humidity shown in Figure 4a-b in the previous section. This section describes the results of a test in which the initial model is used to predict cooling tower water makeup consumption values. This section subsequently explores alternative models.

Out of Sample Prediction

This sub-section uses Model 1 (fitted to data from 2013) to predict values for the 2014 data in an effort to understand how well the model would function if implemented in RTI's building automation system. The results confirm that further analysis is needed.

The challenge associated with examining 2013 and 2014 data with one model is the apparent magnitude of the change between the 2013 and 2014 cooling seasons. In an effort to optimize the predictions, estimated values are multiplied by the scalar that minimizes the overall variance between the 2014 estimates and observed values. Through the Simplex linear programming algorithm, Excel's Solver plugin identified the scalar that minimizes the average absolute value of percent error between the estimated and observed values (4.001813013). Using the scalar, the percent difference for the out of sample prediction is 60.44 percent. Despite the automation of data cleaning to address erroneous negative, outlier and non-explanatory zero values, data problems persist in the initial models.

Based on the results of the out of sample prediction in Table 1-4, Models 1-4 can be improved by transforming variables and considering alternative explanatory variables. The comparison of the full models (Table 1) to the nested models (Table 2) does not

indicate that removing relatively humidity significantly improves the fit of each model, but there is reason to believe that humidity affects demand for chilled water as well as the rate evaporation (and therefore efficacy) of the cooling towers. The primary concern is to find a measure of humidity that has some linear (or near-linear) relationship with the dependent variable. The secondary concern is understanding and addressing the issues with the hourly models.

The R^2 value is important to assessing model fit, but it is only one of a number of criteria that need to be evaluated. Known as the coefficient of determination, R^2 reflects the extent to which the variation in the independent variables explains the change in the dependent variable (Boslaugh, 2012). A high R^2 value is desirable in linear regression models when theoretical justification for the inclusion of explanatory variables exists. R^2 increases in response to correlation between the independent and dependent variables and therefore tends to increase with inclusion of additional independent variables, even if the variables have no theoretical justification for explaining change in the dependent variable (cooling tower water makeup consumption). Relying on models with a high R^2 is important, but R^2 is only a summary statistic. Other measures, including the count of observations (n), statistical significance and correlation between residuals and independent variables (in addition to the correlation between each unique combination of independent variables in models where both exist) are equally important to assessing the usefulness of the linear model.

Subsequent models attempt to improve upon initial models by altering combinations and timing of explanatory variables as well as functional form of the dependent variable.

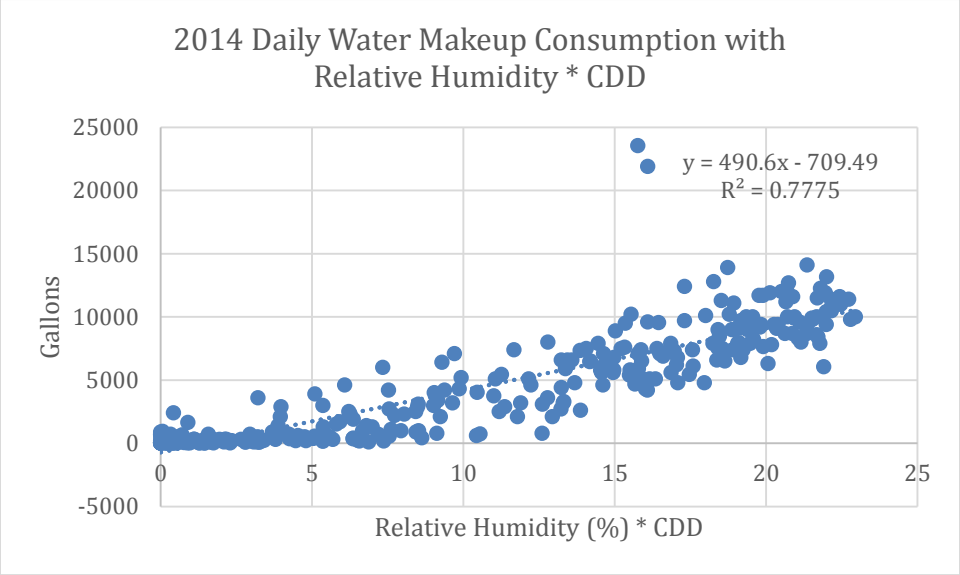
Interactive Variable Models

Temperature and humidity are both essential to predicting the volume of cooling tower water makeup consumption because both are elements of enthalpy that affect the demand for and performance of cooling systems. Although the relationship between humidity and

cooling tower water makeup consumption is not linear, the variation of the combined effect of temperature and humidity may account for the variation in cooling tower water consumption. The interactive variable serves as a proxy for enthalpy, as true enthalpy figures are not present in the data.

A common method of transforming independent variables when the variables have an interactive effect is multiplying the two original variables and interpreting the estimate for the interactive term. Interacting temperature and humidity may better explain the combined effect of temperature and humidity on the demand for chilled water and efficacy of evaporative cooling.

Figure 5: Scatter plot of 2014 Water Makeup Consumption and Relative Humidity * CDD



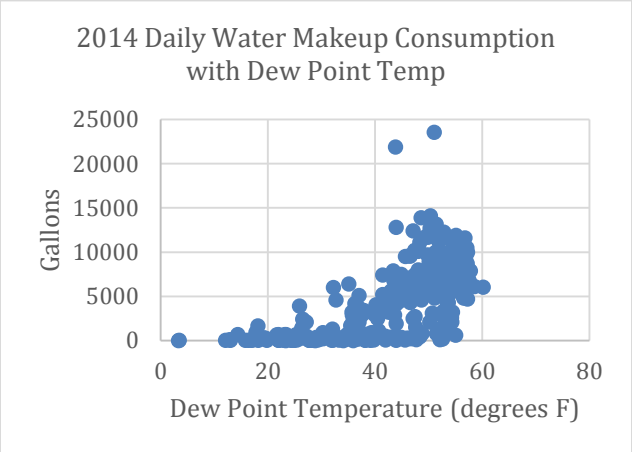
The estimated equations, y = 490.6x - 709.49, indicates a fairly strong linear relationship between the interaction of temperature and humidity and cooling tower water makeup consumption. The R^2 value (77.75 percent) is comparable to the values from the initial model. Most importantly, the correlation (and therefore the estimated coefficient of the variable *x*) is positive, which follows the intuition that cooling tower water consumption increases as temperature and/or humidity increase. The trouble is that increased humidity at a constant temperature may cause water consumption to decrease, since the ability of

liquid water to evaporate depends partially on the saturation of the air surrounding it. Interacting the variables may be useful to consider for the final model, especially if the interaction is based on an alternate measure for humidity that has a clear functional relationship with the dependent variable.

Dew Point and Wet Bulb Temperature Models

Model problems in linear regression can occur due to the distribution, magnitude, or relationships among explanatory variables. Using linear regression to model the relationships between atmospheric conditions and mechanical systems is a challenge due to the nature of the physical relationships that exist among temperature and humidity. Both temperature and humidity are reflected in multiple variables within the NOAA data set. The Excel tool calculates cooling degree days from dry bulb temperature, which reflects only the dry temperature (sensible heat) of the air. Initial models use relative humidity because relative humidity reflects the current absolute humidity relative to the maximum humidity at given temperature and pressure conditions. Relative humidity is therefore independent of temperature and is not high correlated with temperature. Other measures reflect both sensible and latent heat and might be better explanatory variables in the linear model but also might be subject to high correlation if multiple variables are used. This section explores dew point and wet bulb temperature as alternative measures to relative humidity.

Figure 6: Scatter Plot of 2014 Water Makeup Consumption and Dew Point Temperature



The relationship shown is not linear, but is similar to the relationship between cooling tower water makeup consumption and relative humidity depicted in Figure 4a-4b. However, in the full year model (Model 12), the estimated coefficient for dew point temperature is not statistically significant.

In particular, dew point temperature and dry bulb temperature have a high correlation coefficient (0.76). Dew point temperature reflects the outdoor air temperature at which the rates of evaporation and condensation are in equilibrium. Assuming all other factors are constant, the dew point temperature should increase with relative humidity, reflecting the ease in which water condenses from air. But a given value of relative humidity, for example 50 percent, describes much less absolute humidity at an atmospheric dry bulb temperature of 30 degrees Fahrenheit than at 80 degrees Fahrenheit, because colder air holds less water in absolute terms. Due to the fact that dew point temperature is a function of outdoor ambient temperature, it is likely to be highly correlated with temperature and potentially, relative humidity (which can be excluded from the model if dew point temperature is used as an explanatory variable).

Table 4: Results using Dew Point as an Explanatory Measure

D.V. : Cooling Tower Water Makeup Consumption (Gal)

Models use observations where the dependent variable > 0

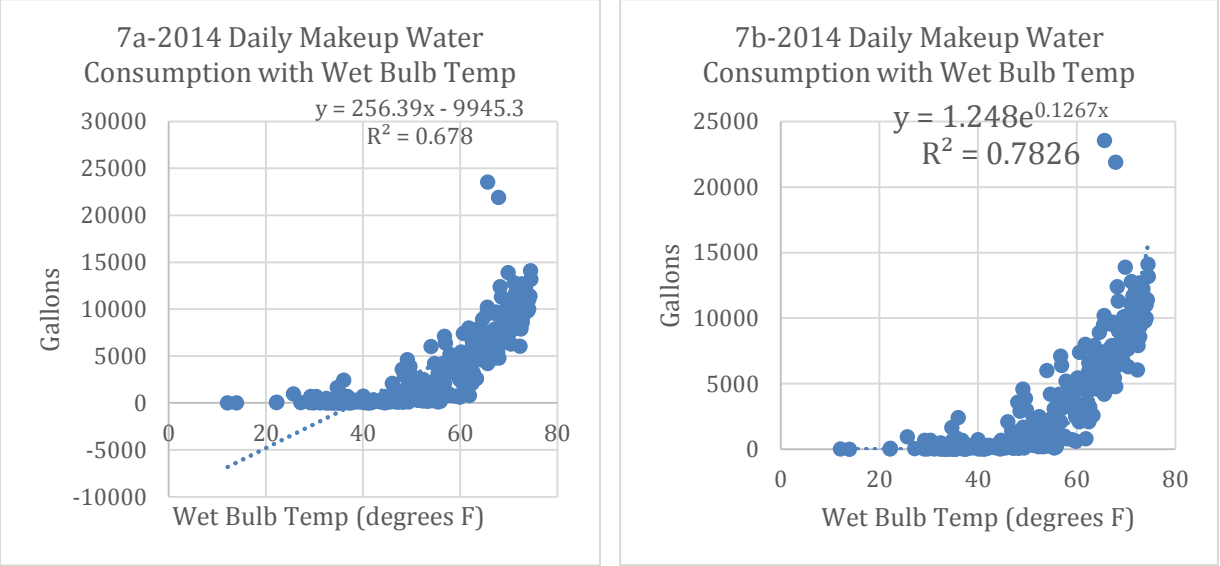
	2014 Dew Point Temp		
	Intercept	DEW Point Temp	CDD
Interval: Daily			
Model 11: Full Year	-797.048	-7.948	367.210
Std Error	447.258	13.830	15.516
Alpha Level		>0.25	0.01
t stat	-1.782	-0.575	23.666
n		311	
R-Square		80%	
Correlation Residuals : D.V.		0.00%	0.00%
Model 12: April 1 - October 31	-3426.576	30.652	401.803
Std Error	854.568	22.272	23.531
Alpha Level		0.10	0.01
t stat	-4.01	1.38	17.08
n		210	
R-Square		72%	
Correlation Residuals : D.V.		0.00%	0.00%

A second alternative to relative humidity is wet bulb temperature, which is the lowest temperature that can be reached by evaporative cooling. Wet bulb temperature yields the same information in the model as dew point temperature and does not yield a better functional relationship than previous models, as is evident by the wet bulb temperature scatter plots (Figure 7) that roughly resemble the dew point scatter plots (Figure 6).

The curvature presented in Figure 7b suggests that the relationship between wet bulb temperature and cooling tower water makeup consumption is curvilinear. When non-

linear trends are present, non-linear models are worth considering. In Figure 7, the slope of the trend line is increasing at an increasing rate, suggesting that an exponential model might better represent the relationship between the variables than a linear model. As Figure 7b illustrates, the exponential models is worth considering, although the trend shown in Figure 7b does not appear to be a significant improvement over previous models.

Figure 7a-b: Scatter Plot of Makeup Water Consumption and Wet Bulb Temperature



Natural Log Transformation Model

Linear regression models most accurately assess data that vary across observations, but vary within a consistent order of magnitude. The cleaned historical data are already void of extreme outliers and negative values, however observations of cooling tower water consumption vary between zero and several thousand gallons. One set of options for improving the fit of the model is using a functional transformation to reduce the magnitude of the dependent variable.

Model 13 uses the natural log of cooling tower water consumption as the dependent variable with dew point temperature and cooling degree days as explanatory variables.

However, the lack of functional relationship between humidity and water makeup consumption is more likely a result of a lack of clear functional relationship than a magnitude issue. Model 13 (Table 5) yields relatively consistent results across full-year and cooling season versions, with statistically significant estimated coefficients for explanatory variables. However, the R^2 values for these models demonstrate that they do not determine variations in cooling tower water makeup consumption significantly better than Models 10 and 11, which also use dew point temperature as the explanatory variable for humidity.

Table 5: Results using the Natural Log of Makeup Consumption as the dependent variable

D.V. : **Ln**(Cooling Tower Water Makeup Consumption (Gal))

Models use observations where the dependent variable > 0

	2014 Dew Point Temp		
	Intercept	DEW Point Temp	CDD
Interval: Daily			
Model 13: Full Year	4.004	0.029	0.138
Std Error	0.214	0.007	0.007
Alpha Level		0.01	0.01
t stat	18.747	4.420	18.640
n		311	
R-Square		79%	
Correlation Residuals : D.V.		0.00%	0.00%
Model 14: April 1 - October 31	5.439	0.021	0.097
Std Error	0.215	0.006	0.006
Alpha Level		0.01	0.01
t stat	25.26	3.70	16.34
n		210	
R-Square		74%	
Correlation Residuals : D.V.		0.00%	0.00%

Lagged Models

Low R^2 values from the hourly models imply that the change is hourly water makeup consumption is not well determined by the hourly changes in explanatory variables. The purpose of building lagged variable models is to examine the possibility that cooling tower water makeup consumption in one period is based on the outdoor ambient temperature in previous periods. The rationale behind lagging the variables is that buildings take some time to absorb changes in atmospheric temperature, resulting in a delay for demand in cooling. The following model includes cooling degree hours as an explanatory variable lagged zero, one, two and three hours.

Table 6: Lagged Variable Model Using 2014 Daily Interval Data

D.V. : Cooling Tower Water Makeup Consumption (Gal)

Models use observations where the dependent variable > 0

	2014 Lagged Temperature				
	Intercept	CDH	CDH_{t-1}	CDH_{t-2}	CDH_{t-3}
Interval: Hourly					
Model 15: Full Year	16.314	8.338	-0.590	2.324	4.017
Std Error	8.733	0.644	0.515	0.510	0.463
Alpha Level		0.01	0.10	0.01	0.01
t stat	1.868	12.949	-1.146	4.555	8.678
n			4467		
R-Square			25%		
Correlation Residuals : D.V.		0.00%	0.00%	0.00%	0.00%

The statistically significant results from the periods lagged zero, two and three hours suggest that some combination of lagged periods contributes to change in cooling tower water makeup consumption. Because the determination (R^2) for this initial model is well below the target 90 percent value, the analysis does not proceed with nested versions.

Multicollinearity

The initial models suffer from a lack of correlation between humidity and cooling tower water makeup consumption. Two similar alternative measures of humidity, including dew point and wet bulb temperature, offer comparable alternatives.

However, both dew point and wet bulb temperature cause multicollinearity in linear models, as demonstrated by correlation tests between independent variables in Table 7. The optimal condition for reducing multicollinearity is including only Dew Point or Wet Bulb Temperature as explanatory variables or reverting to relative humidity and cooling degree days based on dry bulb temperature.

Table 7: Correlation Coefficients among Key Independent Variables

Correlation Coefficients of Independent Variables (2014 Daily Data)

Variable	CDD	Relative Humidity	Dew Point Temp	Wet Bulb Temp
CDD	1.00	0.30	0.76	0.95
Relative Humidity	0.30	1.00	0.83	0.53
Dew Point Temp	0.76	0.83	1.00	0.91
Wet Bulb Temp	0.95	0.53	0.91	1.00

Persistent model limitations include the lack of correlation between atmospheric humidity and cooling tower water makeup consumption, as well as multicollinearity associated with including measures of temperature and humidity. The correlations among explanatory variables shown in Table 7 show that multicollinearity biases multiple regression models that include measures of humidity (aside from relative humidity) as explanatory variables. For this reason, linear regression models cannot include most combinations of the dew point temperature, wet bulb temperature and cooling degree days (or dry bulb temperature).

Conclusions and Recommendations

The greatest challenge is the lack of a linear relationship between measures of humidity and cooling tower water makeup consumption. Unfortunately, functional transformations do not address the limitation for daily models. Future models could investigate the relationship between site enthalpy and cooling tower water makeup consumption. Hourly changes in cooling tower water makeup consumption are partially determined by a weighted combination of temperature changes in previous periods in addition to some measure of humidity. Future studies may investigate the impact of weighting the values of at least several lagged explanatory temperature variables.

Despite robust data acquisition, cleaning and adjustment, major obstacles lie between the available data and a reliable model for real time analysis in RTI's building automation system. Some of the data challenges are related to data quality; others are inherent due to the nature of correlated explanatory variables. Approximate daily assessment is possible using the models explored in this project, but the predicted values will not be reliable for the purposes of comparison. Based on the out of sample prediction, an algorithm built on a linear model with cooling degree days and relative humidity as explanatory variables would have to account for at least 20 percent error associated with the predicted value itself, in addition to the percent error between the predicted value (range) and measured cooling tower water makeup consumption.

As RTI plans system upgrades and capital investments to the building management system hardware and software, the following provisions are worth consideration:

- Collect temperature, humidity and enthalpy data on site.
- Develop the ability to retroactively insert historical data into the repository server to address erroneous data that corresponded to hardware or communication malfunctions.
- Develop advanced machine learning algorithms that utilize methods other than (or in addition to) regression analysis (e.g. principal component analysis).

References

Agnew, K., & Goldberg,, M. (2013, April 1). *Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol.* Retrieved 2015, from energy.gov: http://energy.gov/sites/prod/files/2013/11/f5/53827-8.pdf

BizEE Software Limited. (2014). *Degree Days - Handle with Care!* Retrieved March 2014, from http://www.energylens.com/articles/degree-days

BizEE Software Ltd. (2014). *Linear Regression Analysis of Energy Consumption Data.* Retrieved March 2014, from http://www.degreedays.net/regression-analysis

Boslaugh, S. (2012). *Statistics in a nutshell* (2 ed.). Sebastopol, California: O'Reilly Media, Inc.

Deliso, R. (2013, October 25). *Unpacking Heating Degree Days and Cooling Degree Days.* Retrieved October 3, 2014, from EnergySMART: http://energysmart.enernoc.com/bid/341363/Unpacking-Heating-Degree-Days-and-Cooling-Degree-Days

Johnson, R. A., & Bhattacharyya, K. G. (2010). *Staistics: Principles and Methods* (6 ed.). Hoboken, NJ: John Wiley & Sons, Inc.

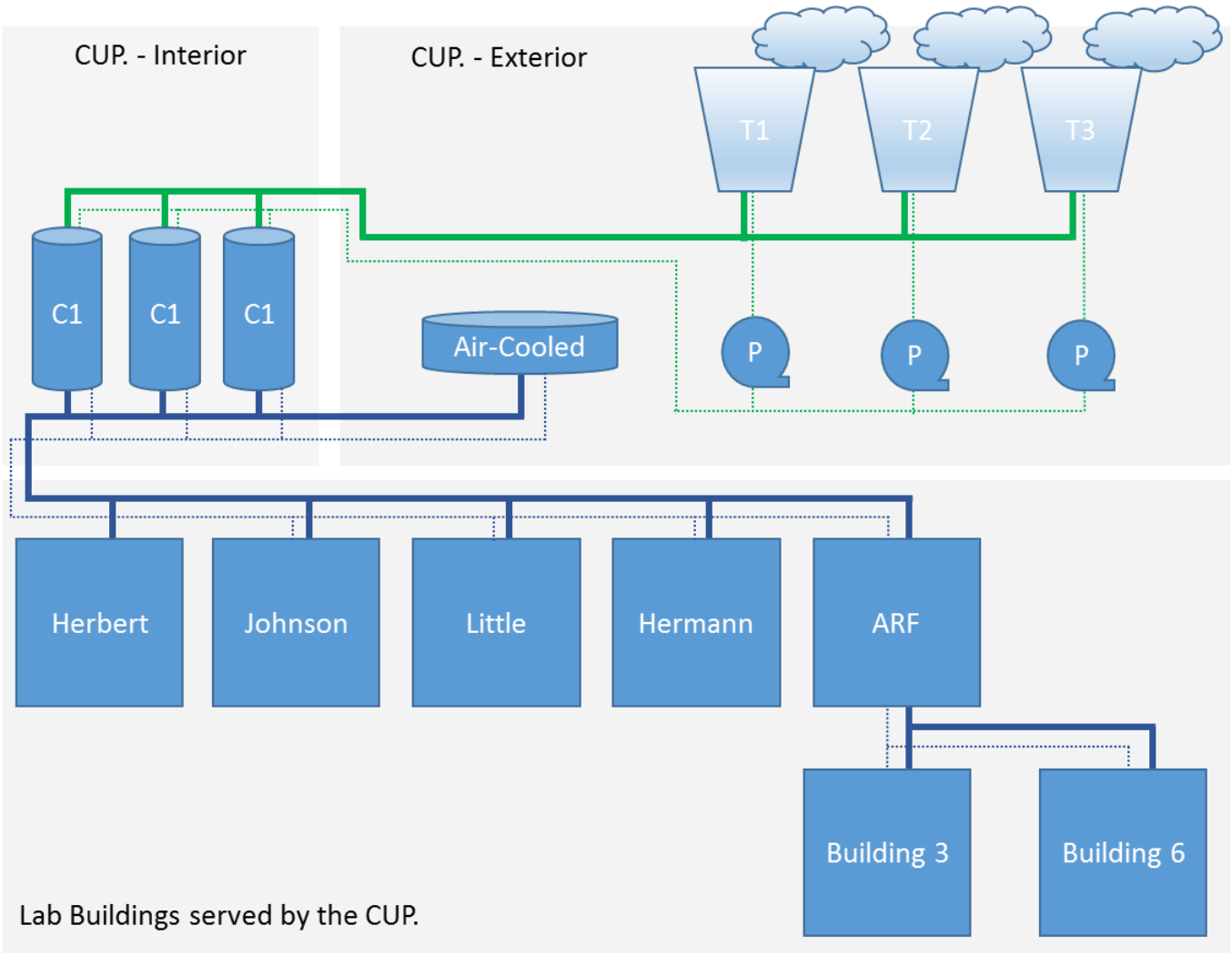
Lam, J. C., Wan, K. K., & Chung, K. L. (2009). An Analysis of Climatic Influences on Chiller Plant Electricity Consumption. *Applied Energy, 86*(6), 933–40. doi:10.1016/j.apenergy.2008.05.016

McMenamin, S. J. (2008). *Defining Normal Weather for Energy and Peak Normalization.* Retrieved from Itron Forecatsing: https://www.itron.com/PublishedContent/Defining%20Normal%20Weather%20for%20Energy%20and%20Peak%20Normalization.pdf

The National Oceanic and Atmospheric Administration (NOAA). (2015). *Quality Controlled Local Climatological Data*. Retrieved from The National Oceanic and Atmospheric Administration: http://cdo.ncdc.noaa.gov/qclcd/QCLCD?prior=N

Appendix

A1 Overview of CUP loop buildings (Cooling Components Only)

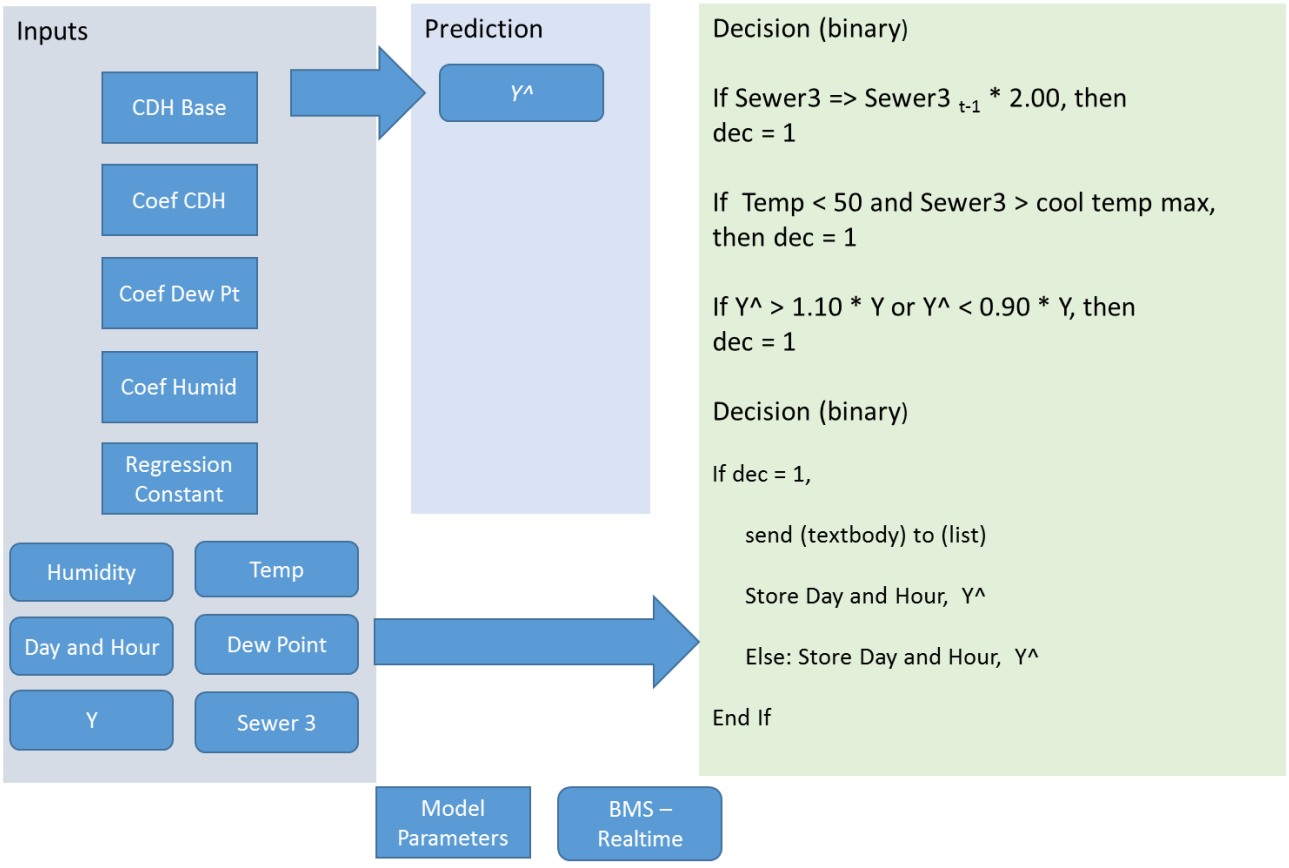


Lab Buildings served by the CUP.

A2 Specifications of CUP Buildings

Building/Space	Year of Occupancy	GSF	% of Total Campus	% of Total RTI USA	Age	End of Life'	Number of Employees *
Building 3	September 1963	10,182	1.12%	0.87%	49	2013	***28***
Building 6	February 1967	10,172	1.12%	0.87%	45	2017	***11***
ARF	June 1971	26,712	2.94%	2.29%	41	2021	***11***
Hermann	June 1971	35,961	3.95%	3.09%	41	2021	***59***
Herbert	March 1984	68,094	7.48%	5.84%	28	2034	***113***
Little	July 1993	72,216	7.94%	6.20%	19	2043	***71***
Central Utility Plant	September 2006	12,115	1.33%	1.04%	6	2051	***0***
TOTAL - RTI CAMPUS		**235,452**	**25.87%**	**20.21%**			***293***

A3 Algorithm Decision Tree



A4 User Guide

Overview

This guide accompanies the macro-enabled Excel workbook completed by Nicholas Garafola, in partial fulfillment of the requirements Master of Environmental Management in Energy and Environment degree in The Nicholas School of the Environment of Duke University

Routine procedures

- Check for errors and missing data
- Clear out existing processed data
- Generate necessary sheets in the workbook if they are not present
- Import BMS (Periscope) from CSV files and NOAA data from user paste
- Handle up to eight consecutive Periscope exports and multiple years of NOAA data
- Assemble BMS and NOAA data into time-series sets (i.e. stack consecutive data intervals)
- Compute region-adjusted timestamps, cooling degree hours, heating degree hours and lagged variables
- Merge BMS and NOAA data sets to form one cohesive data set
- Remove observations that consist of negative or non-numeric temperature and volume data
- Assess outliers based on user discretion and create outlier-removed data set

Assembling Data

Task One: Check Input Assumptions

Review and alter input assumptions on sheet "CONTROL" based on how the weather data appear in Excel ***prior to subsequent steps*** (the macros in Task Three will add additional columns to the left-hand side of each month).

This step should be performed concurrently with Task Two. Most of the inputs do not need to be changed, but check year, row and column references. Below is an example of a valid set of assumptions:

Step 1: Check Assumptions (columns shift in part 3)	
Start Year (first year of data you want to clean)	**2013**
Data starts in row # (default is row 6)	6
Time column # (default is 2 for Col B)	2
Last Column # (usually 23 for W, but sometimes X)	23
Clear first row of each month's data (Type "Yes")	Yes
Month from which to pull master data headers	January
Dry Bulb Temp (F) Column Number (default is 7 for G)	7
CDD/CDH Base (Degrees F, default is 50)	50
HDD/HDH Base (Degrees F, default is 50)	50

Task Two: Paste Each Month's Data from NOAA to Excel

Step One: Open NOAA Data Portal (http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/quality-controlled-local-climatological-data-qclcd)

Step Two: Select "Quality Controlled Local Climatological Data (QCLCD)" from the list of options.

Step Three: Select "North Carolina" from the list of states.

Step Four: Select "***RALEIGH/DURHAM: RALEIGH-DURHAM INTERNATIONAL AP (13722/RDU)" from the list of desired stations. Press "Continue" (underneath the list).

Step Five: Select the first desired year/month from the list and press "Continue" (underneath the list).

Step Six: Click the radio button under "HTML Form" and next to "Hourly 10A." The "E" for "entire month" should be selected by default. Press "Submit."

Step Seven: Using the mouse, highlight the entire table (including the headers), starting with "Date" on the top left and ending with the last value in the right-hand column. Press CTRL and C keys simultaneously to copy the data table. In the Excel file, select the sheet (tab) corresponding to the full name of the month and click on cell A1. On the Excel Ribbon, navigate to the down arrow under "Paste" on the HOME tab. Select the clipboard symbol for "Match Destination Formatting (M)."
For best results, start with January and paste one full year of data into Excel.

Task Three: Stack Data

Once the monthly data sets have been pasted into their respective sheets, press the button labeled "Stack Data" on the left-hand side of the CONTROL sheet. The individual month sheets should disappear while a sheet called "Master" should appear and include the entire year's data. The data set on "Master" include time intervals in the format of the BAS (0:00 – 23:00) but extra intervals will show up as decimal hour values (e.g. 0.98:00).

Task Four: Transform (Calculate Cooling Degree Hours and Clean Data)

Press "Transform" to remove the decimal hour values and compute cooling and heating degree hours. The balance point temperatures (degrees Fahrenheit) can be changed in Task One.

Task Five (Optional): Export Clean NOAA Data

Perform this task to obtain clean NOAA weather data without performing further analysis.
Step One: Press the "Exp CDH" after completing Task Three. A "Save As" Dialog box will appear and will prompt you to select a file path and name.
Step Two: Select the desired location for the weather data output and enter a unique file name in the field adjacent to "File name." Press the button labeled "Save." A dialog box will appear to confirm the location of the new file. The file itself will open upon save.

Task Seven: Import Building Automation System Data

Step One: Prepare hourly-interval .CSV exports of Building Automation System Data, preferably with the following parameters:

Column A: Timestamp
Column B: CUP Cooling Water Blowdown (Gal)
Column C: CUP Cooling Tower Water Makeup (Gal)
Column D: CUP Cooling Condensate Total (Gal)
Column E: Sewer Meter 3 (Gal)

Step Two: Press "Import." Using the prompt, navigate to the folder that contains the .CSV exports. Select the first file and press "Open." The values in the table adjacent to the Import button will automatically update to reflect the import (shown below). The text under "Renamed To" in each row reflects the name of the sheet in the Excel file.

Periscope Files Imported		
File #	**Original File Name & Path**	**Renamed To**
File 1:	C:\Users\Nick Garafola\Dropbox_Masters Project Working Documents\2-Models and Procedures\CUP water and sewer Reports\2014\report Jan 1 2014 - July 1 2014.csv	BASOrig1
File 2:	C:\Users\Nick Garafola\Dropbox_Masters Project Working Documents\2-Models and Procedures\CUP water and sewer Reports\2014\report July 1 2014 - Dec 31 2014.csv	BASOrig2
File 3:		
File 4:		

Step Three: If you would like to import another .CSV, select "Yes" when prompted and repeat Step Two.

Task Eight: Stack BAS Data

Press the button labeled "Stack BAS." This will close the individual BASOrig sheets and create a new sheet called "StackBAS."

Task Nine: Clean Periscope BMS Data

Press the button labeled "Clean Data" to remove outliers and negative values from the BAS data. This will create a new sheet called "BASClean."

Task Ten: Align Periscope and NOAA Data

Press the button labeled "Align Data" to merge the data sets based on Excel time and date serial numbers. This macro pulls the values from the weather data that correspond to the BAS data based on exact matches for time and date, controlling for changes in DST. Duplicate time intervals are deleted.

A5 Alignment Verification

SOURCE	Source Row #	Date	Hour	Makeup (Gal)	CDH
BAS Sheet 1	745	1/31/2014 23:00	23	0	
NOAA	747	1/31/2014	23		0
MERGE	743	1/31/2014 23:00	23	0	0
BAS Sheet 1	2004	3/25/2014 11:00	11	20	
NOAA	2007	3/25/2014	11		0
MERGE	2002	3/25/2014 11:00	11	20	0
BAS Sheet 1	3048	5/7/2014 23:00	23	200	
NOAA	3051	5/7/2014	23		12
MERGE	3045	5/7/2014 23:00	23	200	
BAS Sheet 2	3350	11/17/2014 11:00	11	300	
NOAA	7691	11/17/2014	11		11
MERGE	7684	11/17/2014 11:00	11	300	11

A6 Notes

A6a Dealing with Daylight Saving Time (DST)

The initial version of the Excel Tool (named "Model 6") was used extensively by the client to prepare weather data in support of the 2014 fiscal year sustainability report. Model 6 does not include linear regression functionality but focuses on combining individual months of NOAA data as well as recoding and creating variables.

Model 6 did not successfully account for Daylight Saving Time. A summary of the client feedback follows:

- Gaps in cleaned data can result from missing source data. It might be useful to log a count of missing intervals. John Maravich emphasized a judgment call to address: for duplicate intervals, which interval is the right interval?
- Does the model delete an interval or offset the intervals to account for DST adjustments?

Identifying and communicating missing and duplicate intervals, as with a variety of data cleaning tasks, can be achieved by looping through intervals and/or relying on logical statements.

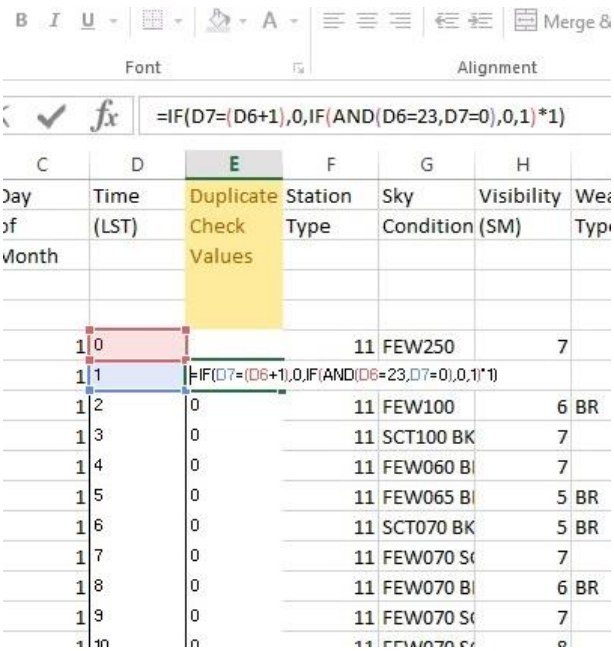
A summary of drafted code that considers a cell in the context of the cell in context of the previous and successive values follows:

```
(Define FirstRow as first row of data observations)
(Define LastRow as last of used range)

Dim Firstval, Secondval as Long
Dim i, j, k as Integer
Define i
'Start at second row because analysis is looking at previous and
post intervals
For j = FirstRow + 1 to LastRow + 1
     Firstval = Sheets("Sheetname).cells((j-1), i).Value
     Secondval = Sheets("Sheetname).cells(j, i).Value

     If Firstval = Secondval Then
          'Take average value of the two intervals FOR ALL
ACTIVE COL
          Sheets("Sheetname").cells((j-1), i).Value = _
          (Firstval + Secondval) / 2
          Secondval = Sheets("Sheetname).cells(j, i).Value = ""
          '(Make the second duplicate value blank for subsequent
deletion.

Next j
```

Alternatively, the solution may be modeled after the client's workaround:

Maravich uses a nested logical statement with the structure: IF(logical_test, [value_if_true], [value_if_false])

The first portion of the logical statement compares the value in D7 to the value of D6+1, which we desire to equal the value in D7. If the comparison passes the first portion of the logical test, the formula returns a value of zero. If the comparison fails the initial test, the formula evaluates the false portion of the logical statement, which itself is a (multicriteria) conditional statement. The second statement checks to see if values that are not successive are equal to the minimum and maximum hour values, which appear as successive values every 24 intervals. Maravich's method results in identification and counting of duplicate values. The loop drafted above can accomplish the same task by storing the number of time each conditional statement is met in one or more variable(s).



External review for DST handling is needed.
Transition to dealing with merging data.

A6b Time Value Comparison

The biggest obstacle in the time use profile analysis is splitting the data based on time and day values. We expect that use patterns are different during times of occupancy and non-occupancy per system setbacks on night and weekend hours. Based on the time stamps, we can incorporate indicator variables into the analysis or split the data to model separate trends for business and non-business hours.

The initial objective is to split data based on weekday and hour variables derived from each time stamp. The Excel worksheet formula "text" provides an easy way to obtain the day of week. To build data sets for business hours and non-business hours, I create columns for each variable that contain conditional statements. The conditional statement for the business hour profile test whether the time interval is Mon-Fri and the time is between 6:00 am and 5:00 pm. In the final model, the beginning and ending values should be user-defined so that the model can adapt to system changes. For the night and weekend ("Non-Business Hour") profile, the conditional statement evaluates whether the interval occurs on a Saturday or Sunday or (even for weekdays), the time is before 6:00 am or after 5:00 pm.

The challenge with creating both formulas is the comparison of numerical time values to the data in the Excel spreadsheet. The initial method was to use the worksheet formula "right" to attempt to extract the hour value from the time stamp column. This method yielded Excel's time and date serial number, which cannot easily be compared to hourly interval counts. The second attempt was to paste the time stamp as a value in a new column with the intent to later automate that step. As with the first attempt, the second

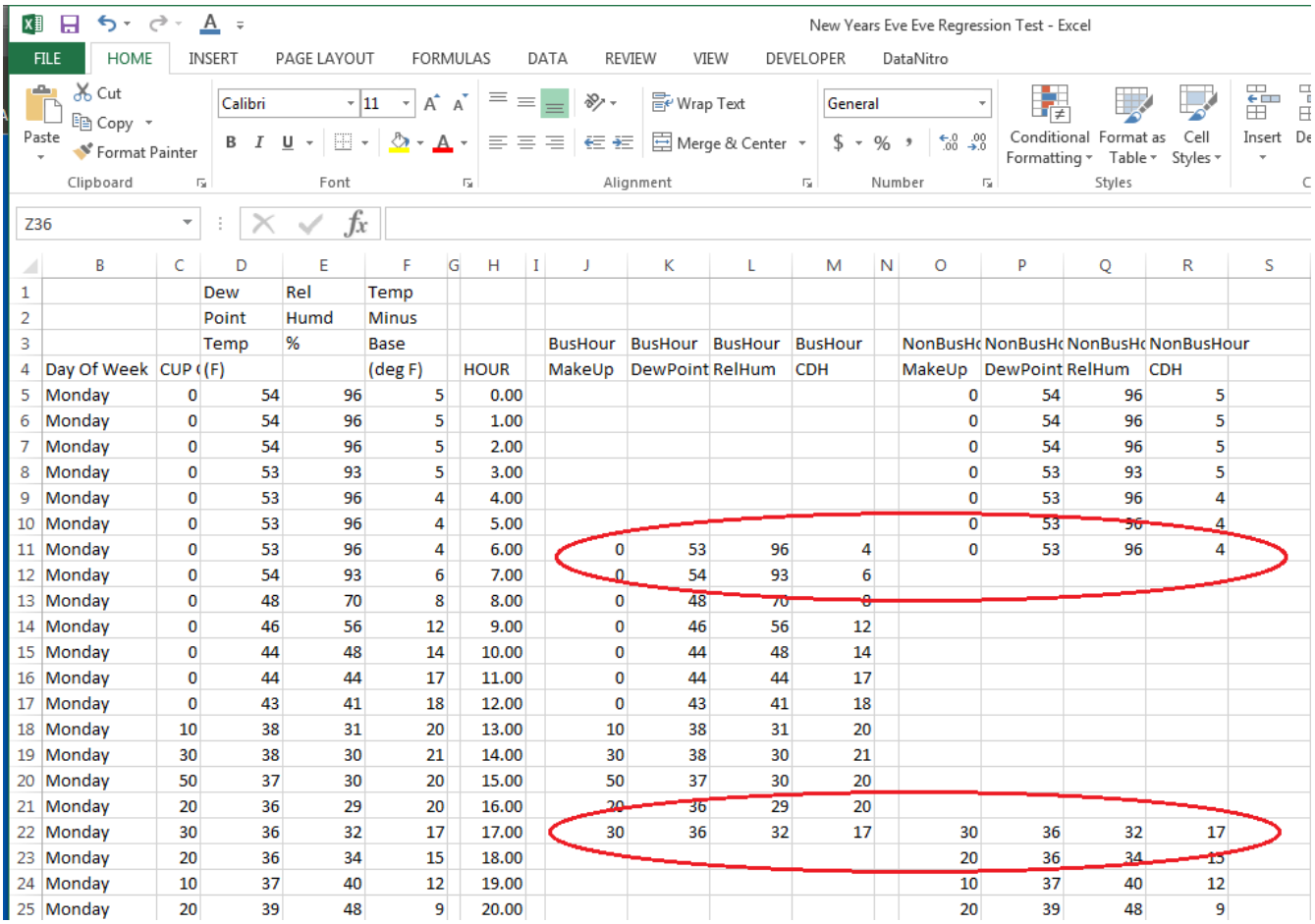
method yielded the time and date serial number, again resulting in an invalid comparison and a blank result from the formula.

The third attempt is a comparison based the Excel "HOUR" formula. The hour formula extracts the numeric hour value (0 – 23) from each time/date stamp. The formulas for the two profiles occur as follows:

Business Hour data point:
=IF((AND(OR($B5="Monday",$B5="Tuesday",$B5="Wednesday",$B5="Thursday",$B5="Friday"),($H5>=6),($H5<=17))),C5,"")

Non-Business Hour data point:
=IF(OR($B5="Saturday",$B5="Sunday",($H5<=6),($H5>=17)),C5,"")

The result of the formulas is values that overlap in both business-hour and non-business hour data sets. For each interval where the hour is equal to 6:00 or 17:00, the formula includes the value in both data sets. The formulas need to be more restrictive. The conditional statements for both profiles cannot include greater-than-or-equal-to statements.

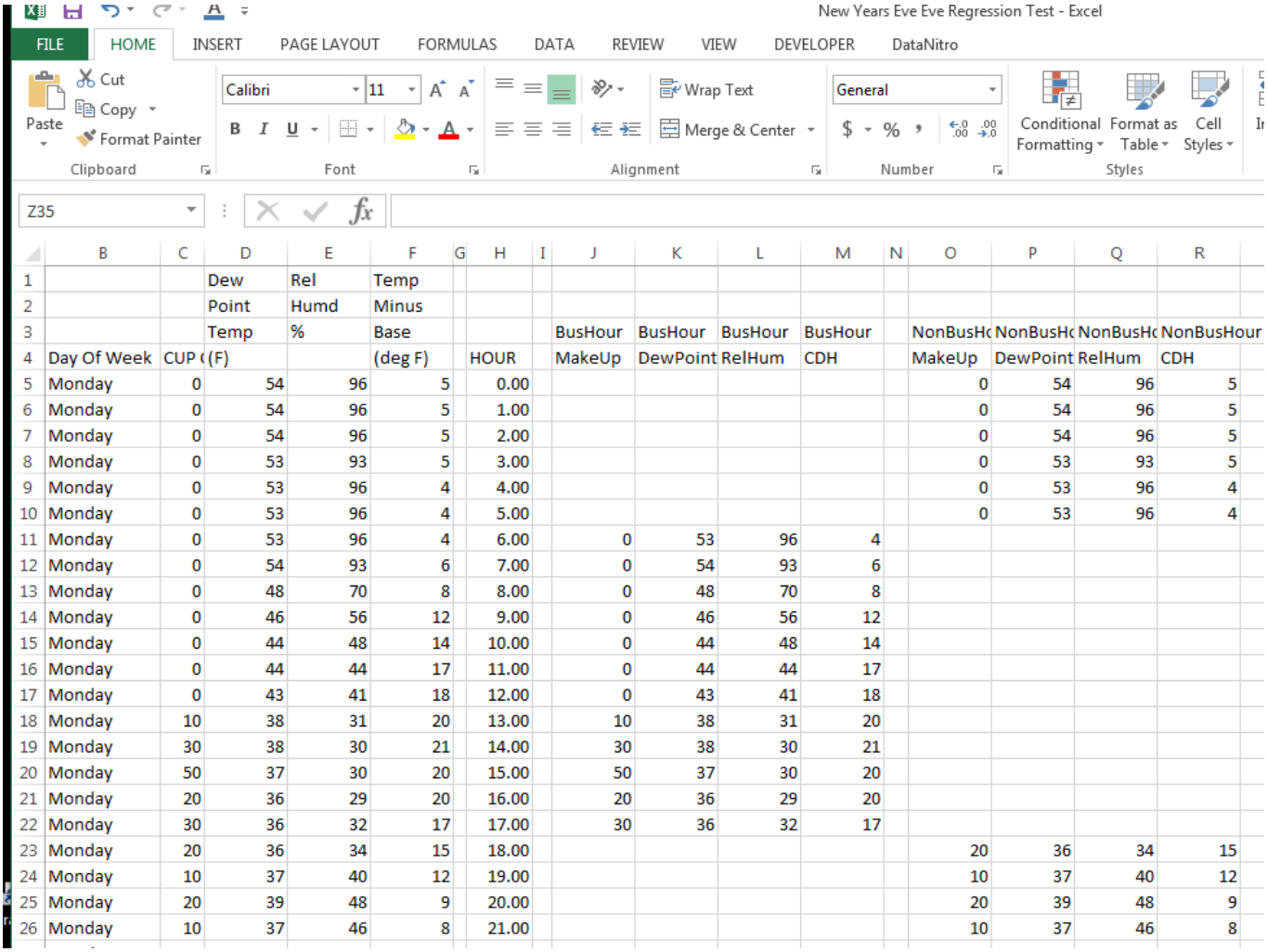


	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1			Dew	Rel	Temp													
2			Point	Humd	Minus													
3			Temp	%	Base				BusHour	BusHour	BusHour	BusHour		NonBusHo	NonBusHo	NonBusHo	NonBusHour	
4	Day Of Week	CUP (	(F)		(deg F)		HOUR		MakeUp	DewPoint	RelHum	CDH		MakeUp	DewPoint	RelHum	CDH	
5	Monday	0	54	96	5		0.00							0	54	96	5	
6	Monday	0	54	96	5		1.00							0	54	96	5	
7	Monday	0	54	96	5		2.00							0	54	96	5	
8	Monday	0	53	93	5		3.00							0	53	93	5	
9	Monday	0	53	96	4		4.00							0	53	96	4	
10	Monday	0	53	96	4		5.00							0	53	96	4	
11	Monday	0	53	96	4		6.00		0	53	96	4		0	53	96	4	
12	Monday	0	54	93	6		7.00		0	54	93	6						
13	Monday	0	48	70	8		8.00		0	48	70	8						
14	Monday	0	46	56	12		9.00		0	46	56	12						
15	Monday	0	44	48	14		10.00		0	44	48	14						
16	Monday	0	44	44	17		11.00		0	44	44	17						
17	Monday	0	43	41	18		12.00		0	43	41	18						
18	Monday	10	38	31	20		13.00		10	38	31	20						
19	Monday	30	38	30	21		14.00		30	38	30	21						
20	Monday	50	37	30	20		15.00		50	37	30	20						
21	Monday	20	36	29	20		16.00		20	36	29	20						
22	Monday	30	36	32	17		17.00		30	36	32	17		30	36	32	17	
23	Monday	20	36	34	15		18.00							20	36	34	15	
24	Monday	10	37	40	12		19.00							10	37	40	12	
25	Monday	20	39	48	9		20.00							20	39	48	9	

The remedy was a modification to the Non-business hour profile formula. I removed the equal-to statements to reflect the following formula:

=IF(OR($B5="Saturday",$B5="Sunday",($H5<6),($H5>17)),C5,"")

The data set has successfully been split into business hour and non-business hour profiles. In the image below, each specific time interval belongs to one profile.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1			Dew	Rel	Temp												
2			Point	Humd	Minus												
3			Temp	%	Base				BusHour	BusHour	BusHour	BusHour		NonBusHo	NonBusHo	NonBusHo	NonBusHour
4	Day Of Week	CUP (	(F)		(deg F)		HOUR		MakeUp	DewPoint	RelHum	CDH		MakeUp	DewPoint	RelHum	CDH
5	Monday	0	54	96	5		0.00							0	54	96	5
6	Monday	0	54	96	5		1.00							0	54	96	5
7	Monday	0	54	96	5		2.00							0	54	96	5
8	Monday	0	53	93	5		3.00							0	53	93	5
9	Monday	0	53	96	4		4.00							0	53	96	4
10	Monday	0	53	96	4		5.00							0	53	96	4
11	Monday	0	53	96	4		6.00		0	53	96	4					
12	Monday	0	54	93	6		7.00		0	54	93	6					
13	Monday	0	48	70	8		8.00		0	48	70	8					
14	Monday	0	46	56	12		9.00		0	46	56	12					
15	Monday	0	44	48	14		10.00		0	44	48	14					
16	Monday	0	44	44	17		11.00		0	44	44	17					
17	Monday	0	43	41	18		12.00		0	43	41	18					
18	Monday	10	38	31	20		13.00		10	38	31	20					
19	Monday	30	38	30	21		14.00		30	38	30	21					
20	Monday	50	37	30	20		15.00		50	37	30	20					
21	Monday	20	36	29	20		16.00		20	36	29	20					
22	Monday	30	36	32	17		17.00		30	36	32	17					
23	Monday	20	36	34	15		18.00							20	36	34	15
24	Monday	10	37	40	12		19.00							10	37	40	12
25	Monday	20	39	48	9		20.00							20	39	48	9
26	Monday	10	37	46	8		21.00							10	37	46	8

A6c Notes: Investigating the Split Data

The final output from the Excel tool utilizes a format similar John Maravich's breakdown of time periods from the post-condensate analysis (late 2014). Maravich's output is more granular than the initial makeup water consumption regression breakdown (section X)

because the data are split into four profiles instead of two. Maravich splits the data into four categories because he expects that the relationship between the independent and dependent variables in each category is distinct based on elements not related to temperature or humidity. For example, occupancy and HVAC setbacks in offices spaces vary across the split profiles.

2014 Full Data Set (January – December)

2014 Cooling Season, no Lag

2014 Cooling Season, Lag

2013 Cooling Season, no Lag

The initial models are based on data split between business and non-business hours, which encompass nights and weekends. The condensate recovery system serves buildings with multiple purposes, while the CUP chillers themselves serve predominately lab spaces. Because lab spaces do not encounter the same occupancy flux or utilize system setbacks, four data profiles are not likely to yield more robust results than two. The downside of four data profiles is that the number of observations N is smaller for each profile, which offers less confidence in the expected value obtained from each model. With a larger number of observations (N>200), we can safely assume that the distribution of the data are normal.

A6d Computing Lagged Variables

Another potential data problem is the delayed effect of weather data on cooling. Gary Bunce's model assessed cooling water makeup and cooling degree days at weekly intervals. For the weekly interval analysis, the delay in impact of temperature and humidity changes on cooling water makeup consumption is likely to be negligible compared the delay associated with an hourly interval analysis. Upcoming models attempt to investigate potential impact of a delay by lagging explanatory variables.

Consider a simplified data set consisting of three observations of time, Y and X (figure 1). The initial regressions associated each Y-value with the X-value that occurs simultaneously (across rows).

Figure 1

$Time_0$	Y_0	X_0
$Time_1$	Y_1	X_1
$Time_2$	Y_2	X_2

However, we suspect that X may have a delayed effect on Y, so we wish to associate our X-value with the Y-value in the subsequent period. We therefore create a new variable equal to X variable lagged by one period. Each value of the new X-variable is computed as follows:

Figure 2

$Time_0$	Y_0	X_0	X_{0-1} = (blank)
$Time_1$	Y_1	X_1	$X_{1-1} = X_0$
$Time_2$	Y_2	X_2	$X_{2-1} = X_1$

The result is the following data:

Figure 3

$Time_0$	Y_0	
$Time_1$	Y_1	X_0
$Time_2$	Y_2	X_1

A6e Computing Lagged Variables Part II

Numerous values for the lagged values of cooling degree hours initially showed up as errors in Excel, despite having valid reference values (i.e. whole numbers) from the CDH lag 0 variable. Troubleshooting the problem, as with most others, required manually stepping through each line of code and following the computation.

In this case, the macro assigns formulas to each of the lagged variable columns that reference the corresponding previous CDH cell.

Upon initially stepping through the macro using year 2014 weather data, the formulas function correctly and populate the lagged CDH variable columns with whole numbers. No errors or missing values are present. A later portion of the Transform macro, which calls CalculateCDHs macro, could be responsible for the errors. But I re-ran the StackData macro, the errors did not appear. The very act of stepping through certain subroutines without following the entire sequence of macros as intended for the user may be the root cause of multiple errors and missing values.

As is common with troubleshooting, the attempt to resolve one error reveals another: the initial formulas that compute lagged variables output zeroes when referencing missing observations.

The downside of assigning formulas to ranges instead of using loops is that logical tests ae limited to Excel worksheet formulas (i.e. any formula the user can type in an Excel worksheet). Multiple logical statements within spreadsheet formulas become quite long and complex and therefore difficult to evaluate and maintain. By contrast, the syntax of logical tests in loops in flexible and generally easy to understand, since the worksheet

values can be stored in named variables and named with concise descriptors, such as "DDBase" for the degree day base value. When multiple logical constraints must be satisfied and computation is not straightforward, calculations within loops in VBA are preferable to assigning formulas to ranges.

I found that when I manually typed formulas into the ranges and referenced the blank cells, the formulas resulted in blank cells. The act of dragging the same formula using relative references down a column range resulted in the zero error. The solution is to copy and paste the range of initial values to the corresponding period of lag instead of using formulas. For example, cooling degree hours lagged by one period use the same exact column as the original cooling degree hours variable, except that the entire column is pasted starting in the second row of data. Using the copy paste function is not only simpler than

The only downside is that extra values will be pasted below the last row. The macro is set to only copy as many values as will fit in the table, based on a final row number equal to the original last row minus the number of periods lagged.